10/717,707 P70-892

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 April 2006 (13.04.2006)

PCT

(10) International Publication Number
WO 2006/038108 A2

(51) International Patent Classification: Not classified

(21) International Application Number:
PCT/IB2005/003001

(22) International Filing Date:
26 September 2005 (26.09.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/615,024 4 October 2004 (04.10.2004) US
60/633,948 7 December 2004 (07.12.2004) US

(71) Applicant *(for all designated States except US)*: **PFIZER PRODUCTS INC.** [US/US]; Eastern Point Road, Groton, CT 06340 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **KIBBEY, Christopher, Edmund** [US/US]; Pfizer Global Research and Development, 2800 Plymouth Road, Ann Arbor, MI 48105 (US). **CALVET, Alain, Pierre** [FR/FR]; 21, Avenue de Carbonelle, F-04860 Pierrevert (FR).

(74) Agents: **FULLER, Grover, F., Jr.** et al.; Pfizer Inc., 201 Tabor Road, Morris Plains, NJ 07950 (US).

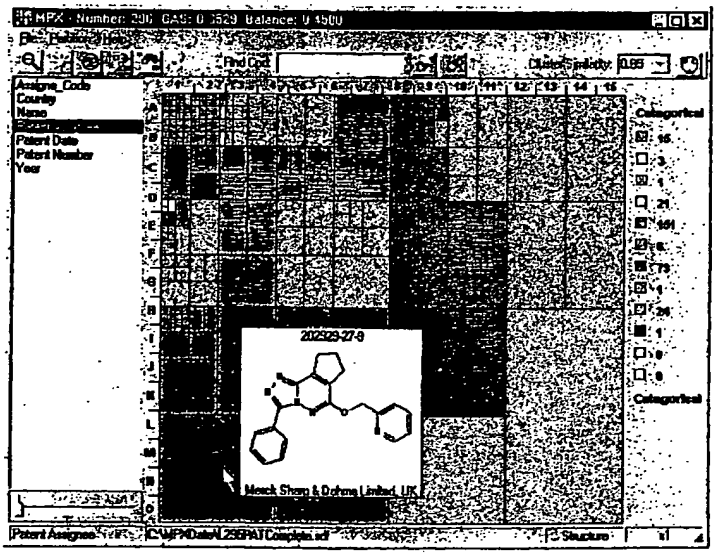
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: METHOD FOR THE READILY COGNITIVE DISPLAY OF STRUCTURE-PROPERTY DATA



(57) Abstract: The method uses agglomerative hierarchical clustering to organize data on the basis of 2D chemical structure similarity across a pre-defined profile of biological assay values and related property values for a plurality of chemical compounds. The resultant hierarchical clusters are visualized as both a treemap and heatmap providing simultaneous representations of cluster members along with their related property values. The heatmap and tree-map are linked and visually integrated. The simultaneous display and integration of the treemap and heatmap provides dynamic readily cognitive, multidimensional visualization of the structure-property data.

WO 2006/038108 A2

METHOD FOR THE READILY COGNITIVE DISPLAY OF STRUCTURE-PROPERTY DATA

This invention relates to the cognitive display of object and related subject data. This invention also relates to the readily cognitive display of structure property data for a plurality of chemical compounds.

BACKGROUND OF THE INVENTION

The substantial increase in chemical structure and biological activity data brought about through combinatorial chemistry and high-throughput screening (HTS) technologies has created the need for sophisticated graphical tools for visualizing and analyzing such data. Visualization of the structure-activity data is important in analyzing and understanding relationships within multi dimensional data sets. The prior art chemoinformatics software applications apply standard clustering techniques to organize chemical compounds on the basis of 2D or 3D structural features, or on the basis of a biological activity associated with the chemical compounds.

The development of software for visualizing multidimensional structure activity data is a significant challenge. Medicinal chemists require that such software be intuitive, provide tools to interact with both chemical structure and biological data, and support the organization and visualization of information-rich data. A number of software tools have been developed during the past approximately ten years to help chemists understand relationships between chemical structure and related activity data.

Navigator, as disclosed in Chapman, D.; Harris, N.; Park, J.; Critchlow, R. E. Jr. Navigator: Tools for informal structure-activity relationship discovery. J. Molecular Graphics 1995, 1 3, 242-249, was developed as a molecular database visualization tool in about 1995. Navigator relies on a maximal common subgraph algorithm to determine neighboring relationships among chemical structures. This approach to data organization is intuitive to most chemists, because it facilitates comparisons between compounds. Two compounds are considered related if more than half their structure is identical and if one molecule can be transformed into the other by breaking a single bond and replacing the substituent at this position.

VisualiSAR, as disclosed in Wild, D. J.; Blankley, C. J. VisualiSAR: A Webbased application for clustering, structure browsing, and structure-activity relationship study. J. Molecular Graphics 1999, 17, 85-89, is a web-based program that employs modal fingerprints along with Stigmata coloring of atoms to highlight common and unique structural features among compounds at various levels wiihin a hierarchical cluster. VisualiSAR employs Daylight fingerprints, as disclosed in Cosgrove, D. A.; Willett, P. SLASH: A program for analyzing the functional groups in molecules. J. Molecular Graphics 1998, 16, 19-32, and Ward's, as disclosed in Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. J. Amer. Stat. Assoc. 1963, 58, 236-244, disclose clustering to organize chemical structures on the basis of their Tanimoto similarity. In addition, that software provides navigation tools for selecting among the various levels of the cluster hierarchy and displaying the chemical structures of cluster members. While VisualiSAR provides a useful means for visualizing chemical structures within specific clusters and cluster levels, it suffers from a common problem associated with prior art visual representation and navigation of hierarchical data. That is, VisualiSAR does not adequately convey the spatial relationships between clusters and cluster members among the various levels of the hierarchy.

An alternative to hierarchical agglomerative clustering is Optimizable KDissimilarity Selection (OptiSim), as disclosed in Clark, R. D. OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. J. Chem. Inf. Comput. Sci. 1997, 37, 1181-1 188, which Tripos, Inc. employs the SARNavigator, as disclosed in SARNavigator is available from Tripos, Inc. http:llwww.tripos.coml, product. SARNavigator presents structureactivity data in a "landscape view," wherein structurally similar compounds or clusters are plotted as circles of varying size within the central region of the landscape and dissimilar compounds are placed along the perimeter. The size of a circle represents the boundary of the cluster in structure space, and the circles are colored to correspond with specific activity data. While the "landscape view" is effective at providing a unifying visualization of structure and activity data, the relationship between compounds plotted in the central region of the landscape and those along the perimeter is effectively lost.

An alternative approach to analyzing structure-activity data focuses on identifying substructures that correlate with activity. The program SLASH, as disclosed in Cosgrove, D. A,; Willett, P. SLASH: A program for analyzing the functional groups in molecules. J. Molecular Graphics 1998, 1 6, 19-32, generates a set of functional groups from an input file, and then analyzes the distribution of these groups among the active compounds in the input data. The program LeadScope, as disclosed in Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E. Jr. LeadScope: Software for Exploring Large Sets of Screening Data. J. Chem. Inf. Comput. Sci. 2000,40, 1302-1 314, keeps track of the number of compounds that possess specific functional groups, aromatics and heterocycles. Users of SLASH may exclude structures from consideration by setting limits on the range of specific structure (e.g., molecular weight, logP, and number of rotatable bonds) and activity data. LeadScope relies heavily on the use of histograms, and scatter-plots, neither of which are well suited to visualizing SAR.

The challenge in presenting multidimensional, chemical structure-activity data lies in the mapping of this data onto a two- or three-dimensional space. A common approach to reducing the dimensionality of a data set is non-linear mapping, and this is usually achieved through principal-component analysis (PCA), or multidimensional scaling (MDS). An alternative to PCA and MDS is the use of Kohonen neural networks to construct Self-organizing Maps, as disclosed in Kohonen, T. In Self-Organizing Maps, 2"d Edition, Springer-Verlag, Berlin, 1997. Gedeck and Willett, as disclosed in Gedeck, P.; Willett, P. Visual and computational analysis of structure-activity relationships in high-throughput screening data. Curr. Opin. Chem. Bio. 2001, 5(4), 389-395, elaborates on the application of these techniques to visualizing structure-activity relationships (SARS) in high-throughput screening data in a recent review article. When applying non-linear mapping to structure-activity data, there is always a trade-off between choosing structure, or activity as the primary means of representing the data. Organizing structure-activity data on the basis of chemical structure often interferes with the presentation of the corresponding activity data. Similarly, multidimensional activity data represented in 2D, or 3D plots makes it difficult for a chemist to grasp underlying correlations between chemical structure and activity. Data visualization programs, such as SpotFire, as disclosed in SpotFire Decision site is available from SpotFire, Inc. http://www.spotfire.com/, are effective in their approach to representing multidimensional activity data in two or three dimensions. However, SpotFire does not manage chemical structures natively, nor does it provide dynamic visualization of hierarchical clusters.

Hierarchical clusters generally are represented as dendrograms, which can be difficult to navigate, especially when applied to large data sets. In about 1992, B. Shneiderman, as disclosed in Shneiderman, B. Tree visualization with Tree-maps: a 2-d space-filling approach ACM Trans. Graphics 1 992, 1 1, 92-99, created tree-maps as an alternate visualization for large, hierarchical data structures. A tree-map is a 2D space-filling approach in which each leaf of a tree is represented as a rectangle whose size and fill color correspond with specific attributes in the data being represented. The tree-map algorithm is also applied to depict hierarchical computer file systems in an efficient manner. Recently, the tree-map algorithm was applied to aid the visualization of gene expression data within the context of the gene ontology, as disclosed in Baehrecke, E. H.; Dang, N.; Babaria, K.; Shneiderman B. Visualization and analysis of microarray and gene ontology data with treemaps. BMC Bioinformatics 2004, 5(1), 84-96.

A heatmap is used to visualize gene expression data of certain drug methodology enzymes in rat livers after treatment, as disclosed in Naoki Kiyosawa, Toshiyu ki Watanabe, Kyoko Sakuma, Miyuki Kanbori, Fu kuroi, Shizuoka, Ltd., Japan, Toxicology Letters 2003, 145(3), 281-289.

Data mining is generally understood to be a process that uses computerized data analysis tools to provide data patterns and relationships that are useful to draw conclusions. The objective of data mining is to produce, from given data, some new knowledge or insight. Data mining generally relies on a large number of databases, with resultant large numbers of indices and files. Such large numbers of indices and files are difficult to view and navigate, and are not readily visually cognizable.

The diverse arts including, by way of examples, the data mining and chemoinformatics arts, desire a dynamic and readily cognitive visualization of large volumes data. The chemoinformatics art desires a method and system for the readily cognitive multidimensional visualization of data, particularly including structure and related property data for chemical compounds. The present invention provides a solution to these diverse arts desired needs.

SUMMARY OF THE INVENTION

The present invention is a visualization tool for hierarchically structurable data, which provides a readily cognitive display of the data. In one principal aspect, the present invention is a method for the readily cognitive display of data for a plurality of subjects and their related object, which method includes: displaying the data in a tree-map, displaying the data in a heatmap, and integrating the tree-map and heatmap, whereby there is a readily cognitive display of the data. The invention employs a reciprocal nearest neighbor (RNN) algorithm to organize the data into hierarchical clusters and sub-clusters.

In another principal aspect, the present invention is a method for the readily cognitive visual display of the structure-property data of a plurality of compounds. In this aspect of the invention the method includes: displaying the structure-property data in a tree-map, displaying the structure-property data in a heatmap, and integrating the tree-map and heatmap, whereby there is a readily cognitive multidimensional display of the chemical structure-property data.

In yet another aspect, the present invention is a method for the readily cognitive display of the structure-property data of a plurality of compounds which method includes, organizing the structure-property data by agglomerative hierarchical clustering, and displaying the structure-property data in a tree-map, wherein there is a readily cognitive multidimensional display of the structureproperty data.

In still another aspect, the present invention a method for the readily cognitive display of the structure-property data of a plurality of compounds comprising: organizing the structure-property data by agglomerative hierarchical clustering; and displaying the structure-property data in a heatmap; whereby there is a readily cognitive multidimensional display of the structure-property data.

The tree-map has a plurality of defined regions (e.g. rectangles), and the heatmap has a plurality of cells. Each heatmap cell is a respective row and column intersection. There is a 1 : 1 correspondence between a specific heatmap cell and a specific tree-map region. The tree-map and heatmap are integrated, whereby the user readily mouse links and navigates between the heatmap and tree-map displays.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1A is a representation of a hierarchical cluster of ten compounds shown as a tree-map.

Figure 1B is a representation of a hierarchical cluster of ten compounds shown as a heatmap.

Figure 2 is a tree-map of 1883 compounds from the NCI GIs0 diversity data set, with the compounds clustered according to the similarity between their 2D chemical fingerprints, and the rectangles of the tree-map colored according to each compound's GIs0 in the OVCAR- 3 cell line.

Figure 3 is a tree-map of 1883 compounds from the NCI diversity data set, with the compounds clustered on the basis of their GIs0 profile across the OVCAR-3, -4, -5, and -8 cell lines, and the rectangles of the tree-map shaded according to each compound's in the OVCAR-3 cell line.

Figure 4 is heatmap representation of the clustered profile from Figure 3, with the compounds with high values across the OVCAR-3, -4, -5, and -8 cell lines visible in the upper half of the heatmap, and wherein a portion of this region of the heatmap has been selected, with the corresponding structures for these compounds shown in the dialog below the heatmap.

Figure 5 is a heatmap of the NCTRER estrogen receptor binding data set, showing the compounds in the data set clustered according to the similarity between their 20 chemical fingerprints, wherein a relationship between structure and activity is readily cooperatively apparent from the overlap of activity category and assigned chemical class for the most active estrogen receptor binding compounds in the data set.

Figure 6 is a tree-map of the most active estrogen receptor binding compounds in the NCTRER data set showing the compounds clustered according to the similarity between their 2D chemical fingerprints, wherein the compounds tend to cluster according to their assigned chemical class as illustrated by the annotations added to the tree-map.

Figure 7A shows the structures of genistein and three related phytoestrogen isoflavones of Figure 6.

Figure 7B shows the structures of four steroids possessing aromatic A-rings selected from the tree-map of Figure 6.

Figure 8 is a Schneiderman prior art of the entire fly genome in the context of the gene ontology.

Figure 9 is a selection of the enzyme-activity region of the molecular-function region from the tree-map of Figure 8.

Figure 10 is a selection of the peptidase-activity region from Figure 9 showing which genes of the fly genome are associated with various peptidase enzymes, wherein the size and color of the rectangles displayed are based on the parameters set in the legend panel to the right of the tree-map.

Figure 11 is a tree-map of 296 compounds obtained from a patent search of GABA-A a5 agonists, which tree-map shows a clustering of the compounds based on the similarity of their chemical structures, and wherein the colors of the rectangles in the tree-map represent the respective assignees of the corresponding patented compound.

DESCRIPTION OF THE INVENTION

Definitions

The term "property used hereinbefore and hereinafter means a physical parameter, chemical parameter, physico-chemical parameter, biological parameter, clinical information, patent information, and other information, as well as biological activity (the term "activity" being afforded its well-understood meaning in the art). The following without limitation are examples of a property within the meaning of the present invention: biological (e.g., %inhibition, %activity, IC-50), chemical class, therapeutic target, physicochemical (e.g., log P, log D, pKa, polar surface area, dipole moment), patent information (e.g., patent assignee, patent issue date, inventor), and clinical information (e.g. dose, indication, side effect data).

The term "patent" as used herein in connection with describing and claiming the invention broadly contemplates and means e.g. issued patents, published patent applications, statutory invention requests, disclosure documents, interference filings, reexamination filings, protests, defensive publications, and like U.S. and foreign documents.

The present invention, however, in addition to patents as defined hereinabove, broadly contemplates any information or informational materials, and by way of example without limitation includes: published or publicly available information (e.g. scientific, technical, medical, chemical and pharmaceutical journals, articles therein and abstracts of the articles), legal publications (e.g. case law, law journals), not generally speaking publicly available information such as company or business documents, (e.g. research reports, correspondence) and the like.

The term "SAR" as used herein is a well-understood term in the chemoinformatics art and refers to chemical structure-activity relationship.

The term "SPR" as used herein is a coined term that refers to structure propertyrelationship, wherein the term "property" is as previously defined.

The term "recursive" as used herein is a coined term derived from the recursion process of defining an object in terms of itself, as generally discussed by Kenneth H. Rosen, in "Discrete Mathematics and its Applications," 4th Ed. 1999, pp. 202-219.

Tree-Map and Heatmap

The tree-map algorithm traverses the branches of a hierarchical cluster recursively beginning with the root node. The tree-map algorithm begins with a defined region (rectangle) corresponding to the root of the hierarchical cluster. As each branch is visited, a rectangular region of the tree-map is split evenly along alternating vertical and horizontal centers. Upon reaching a terminal node in the hierarchical cluster, the corresponding rectangular region of the tree-map is associated with the terminal node of the cluster. Consequently, each rectangular region of the tree-map is uniquely associated with a single node in the cluster. An example of a tree-map generated from a hierarchical cluster of ten objects is shown in Figure 1 (A). The perimeter of the tree-map corresponds to the root of the hierarchical cluster shown on the left side of Figure 1 (A). The root node is split into a left and right

branch, and this is reflected in the tree-map by dividing the rectangular region vertically into two equal halves. The left half of the treemap corresponds to the left branch from the root of the cluster, and the right half of the tree-map corresponds to the right branch. The left branch of the cluster also possesses a left and right branch. The left half of the tree-map is split horizontally into two equal halves. The upper half represents the left sub-branch, while the lower half corresponds to the right sub-branch. Continuing with the left sub-branch in the cluster, the upper quarter of the tree-map is split vertically, and the left half of this rectangular region is assigned to node 1 in the cluster. Traversing the remaining branches of the hierarchical cluster results in the treemap shown on the right side of Figure 1 (A). Each rectangular region of the treemap is identified by a distinguishing color corresponding to the value of a property (e.g., biological assay) associated with the chemical structure of the corresponding node in the cluster hierarchy. The following pseudo-code illustrates one possible implementation of the tree-map algorithm.

A Node structure retains information related to agglomerative hierarchical clustering (compoundID, leftchild, and rightchild) and the coordinates of its rectangular region in the tree-map.

```
struct Node
int compoundID;
Rect rectang ularReg ion;
Node leftchild;
Node rightchild;
void CreateTreemap(N0de node, Int left, int top, int right, int bottom, boo1
axis)
If (node.leftChild = = 0 and node.rightChild = = 0)
//this is a terminal sub-cluster
node.rectangularRegion = Rect(left, top, right, bottom);
else if (axis)
//Divide rectangular legion vertically
CreateTreemap(node.leftChild, left, top, left+(right-left)/2, bottom, !axis);
CreateTreemap(node.rightChild, left+(right-left)/2, top, right, bottom, !axis);
else
//Divide rectangular region horizontally
CreateTreemap(node.leftChild, left, top, right, top+(bottom-top)/2,
!axis);
CreateTreemap(node.rightChild, left, top+(bottom-top)l2, right,
bottom, !axis);
```

The command "CreateTreemap(root, 0, 0, 200, 200, true)" creates a treemap 200 units wide by 200 units high, and begins with a vertical division of this bounded region.

In the aforesaid manner, the tree-map provides immediate insight into the spatial relationship between items and among sub-clusters of the hierarchical cluster. The size of a rectangle in the tree-map correlates with the depth of the corresponding node in the hierarchical cluster. Compounds present at the same cluster level are depicted as rectangular regions of equal size in the tree-map. Compounds within a sub-cluster are depicted by a smaller tree-map bounded by a rectangular region

within the main tree-map. A clustered set of structurally related and diverse compounds results in a tree-map characterized by regions of densely packed rectangles interspersed with more sparse regions. The sparse regions of a tree-map correspond to compounds belonging to sub-clusters that lie closer to the root of the hierarchical cluster.

A heatmap depicts a hierarchical cluster of items along its y-axis together with a hierarchical clustering of property data along its x-axis. An example of a heatmap representation of a cluster of ten compounds across five properties is shown in Figure 1 (B). The left most terminal node in the cluster of compounds corresponds to the first row of the heatmap, and the right most terminal node corresponds to the last row. Likewise, the left most item in the cluster of properties corresponds to the first column of the heatmap, and the right most item in the cluster corresponds to the last column. At each row-column intersection, a rectangle is drawn and shaded to represent the value corresponding to that particular compound-property pair. Each rectangle of the heatmap is the same size.

It has been found that a heatmap and corresponding tree-map provide complementary visualizations of clustered structure-activity data. The tree-map conveys both the topology of the corresponding hierarchical cluster and secondary information associated with each item of the cluster. In addition, the space-filling characteristics of the tree-map algorithm enable every sub-cluster within a dendrogram to be viewed simultaneously. For example, there are four sub-clusters at a depth of 2 from the root in the dendrogram of Figure 1 (A). These four clusters are represented by the four quadrants of the corresponding tree-map: the upper-left quadrant is cluster (1, 2, 3); the lower-left quadrant is cluster (4, 5); the upper-right quadrant is cluster (6, 7, 8, 9); and the lower-right quadrant is cluster (10). A tree-map is limited to depicting only one property associated with items of the cluster. In contrast, a heatmap provides a visualization of cluster nodes across multiple property values. A heatmap, however, does not depict the hierarchy that exists between nodes within the cluster. When applied to a common hierarchical cluster of data, a tree-map may be regarded as a more detailed representation of a columnar cross-section of a heatmap.

MPX System

The present system is otherwise referred to herein as "MOLECULAR PROPERTY EXPLORER" or "MPX". The MPX graphical user interface consists of four major components (see e.g. Figure 2). The menu bar provides access to commands for opening a data set, modifying the graphical representation of the data, partitioning the data into smaller subsets on the basis of property or chemical structure criteria, and accessing on-line help. Below the menu bar is a tool bar that contains buttons for scaling the display region, toggling between heatmap and tree-map visualizations, clustering the data set, adding compounds to the data set, searching for compounds by name or by structure, and cycling through the display of property data over a predefined animation interval. The heatmap/tree-map display region occupies most of the application window. The tree-map visualization is annotated with a virtual map grid with letters along the left and numbers along the top edge of the tree-map. This grid provides visual reference to the absolute location of a zoomed region of the tree-map. To the right of the tree-map is a legend that defines the color assigned to each rectangle of the tree-map over a linear range for the selected property. A list of properties read from the data set is displayed to the left of the display region. Single or multiple items may be selected from the list, and the software automatically updates the display region to reflect the currently selected property. When multiple

properties are selected from the property list, the slider below this list becomes active and may be used to choose amongst the set of selected properties (Figure 2).

Chemical structures must first be organized into a hierarchical cluster prior to visualization as a tree-map, or heatmap. MPX clusters a data set on the basis of 2D chemical structure, or a set of related properties defining a profile. When clustering by chemical structure, MPX relies on the Accord Chemistry SDK to generate 2D fingerprints from chemical structures. The Accord Chemistry SDK uses an approach similar to the Daylight method of computing fingerprints from 2D structures (as previously discussed herein). The Accord Chemistry SDK is disclosed in The Accord SDK is available from Accelrys, Inc. http:llwww.accelrys.coml. MPX uses this fingerprint data to populate a lower triangular matrix with the Tanimoto similarity between pairs of compounds in the data set. Tanimoto similarity between a pair of chemical structures (A, B) is defined by Equation 1 :

$$CTAB = a+b-c(1)$$

where a and b are the number of 1 bits appearing in the fingerprint of structures A and B, respectively and c is the number of 1 bits in common to the fingerprints of structures A and B. The MPX software uses matrix of Tanimoto similarity to cluster compounds, compute centroids of sub-clusters, and compute a group average similarity (see below) for the data set.

Clustering is achieved using a reciprocal nearest neighbors (RNN) algorithm, and consists of two primary steps: computation of the distance between all items in the data set, followed by an agglomeration process in which sub-cluster hierarchies are formed. A discussion of the reciprocal nearest neighbor algorithm (RNN) is found in Murtagh, F. A Survey of Recent Advances in Hierarchical Clustering Algorithms. Computer Journal 1983, 26(4), 354-359. When clustering by chemical structure similarity, distances between pairs of compounds are computed as l-Tamimoto. When clustering by property profile, the distance between pairs of data may be computed in one of four ways: Canberra, cosine, Euclidean, and l-Tanimoto. Five linkage algorithms are supported within MPX: single, complete, un-weighted arithmetic average, weighted arithmetic average, and Ward's, (as previously discussed herein).

The MPX software displays three metrics in the title of the application window to aid interpretation of the corresponding hierarchical cluster. These are the number of compounds in the cluster, the group average similarity (GAS) between compounds, and the balance of the hierarchical cluster. The group average similarity is computed as the mean of the average Tanimoto similarity across rows of the similarity matrix, ignoring self-similarity. GAS ranges from 0 to 1. Data sets consisting of diverse chemical structures produce a GAS of approximately 0.5, while more focused data sets (e.g., combinatorial libraries) yield a GAS between 0.80 - 0.85.

Balance is a measure of the depth of a hierarchical cluster relative to the minimum depth of a binary tree consisting of equal capacity. Equation 2 defines balance of a hierarchical cluster:

$$\text{ceil}(\log, (N)) \text{ Balance } = D$$

where ceil is the ceiling function, N is the number of compounds in the cluster, and D is the actual depth of the hierarchical cluster. The numerator in Equation 2 above defines the minimum depth of a binary tree consisting of N nodes. Consequently, a balance of 1.0 indicates a hierarchical cluster with minimum depth, whereas a balance close to zero describes a hierarchical cluster with excessively long branches.

Each rectangular region of the heatmap/tree-map represents a single compound in the data set. Clicking and dragging with the left mouse button over a region of the heatmap or tree-map displays the corresponding 2D structures within a separate Structure Viewer window. Selected regions of the heatmap/tree-map are outlined in black. Clicking on a selected rectangle will outline it in red and highlight its corresponding structure in the Structure Viewer with a white background. A user may export selected structure-activity data in SD or tab-delimited text format from the Structure Viewer dialog.

The MPX software utilizes all of the information contained within a structure-property data set at once. However, one may be interested in visualizing specific subsets independent of the larger data set. The MPX software provides four means of partitioning a data set into smaller subsets for independent visualization. The data in such subsets are represented as heatmaps/tree-maps in dialog windows separate from the main application window. A data set may be partitioned on the basis of property, or 2D structure criteria. Partitioning on the basis of property criteria involves specifying discrete ranges for a set of properties. Only compounds whose property values lie within each specified range are included in the subset.

MPX offers three methods for partitioning a data set on the basis of 2D chemical structure criteria: substructure match, similarity match, and R-group analysis. The subsets generated from such partitions provide insight into the influence of specific structural features on SAR. The software partitions a data set on the basis of substructure criteria by identifying compounds in the data set that contain specified substructure(s). The user may define multiple substructure criteria and specify whether a matching compound must contain all substructures, or at least one substructure. A partition based on chemical similarity identifies those compounds of the data set that meet or exceed a minimum Tanimoto similarity to a set of query compounds. Lastly, the MPX software can perform an "R-group" analysis of a set of compounds possessing a common core structure. The user draws a core substructure with designated "Rgroups", and the MPX software generates lists of the unique substituents present at each attachment point for all compounds in the data set that possess the core substructure. These substituent lists are then used to define a query structure to be used in a subsequent substructure search of the data set. Compounds that match the query structure are included in the partitioned subset.

Two examples illustrate the use of the MPX method to visualize multidimensional structure-activity data sets. The first example employs the GIs0 diversity set obtained from the National Cancer Institute's Developmental Therapeutics Program (see Screening data and 20 structures of compounds in the NCI GI50 diversity data set and available online: http://dtp.nci.nih.gov/webdata.html). The assay values in this data set are reported as the negative log of the concentration of compound required to inhibit the growth of a tumor cell line by fifty percent. The second example consists of an estrogenic receptor binding data set obtained from the National Center for Toxicological Research Estrogen Receptor (NCTRER) binding database, as discussed in Structures and estrogenic receptor binding data are available online: http://www.epa.gov/nheerl/dsstox/sdf-nctrer.html.

The tree-map of Figure 2 represents a clustering by 2D chemical structure of 1883 compounds in the National Cancer Institute's diversity data set. The group average similarity and balance of the hierarchical cluster are shown in the title bar of the MPX application window. The group average similarity for this set of structures is 0.5280, and this value is consistent with a set of structurally diverse

compounds. The balance of the hierarchical cluster of the compounds is 0.2037, and suggests the underlying hierarchical cluster possess branches of considerable depth. Indeed, this is apparent from the regions of densely packed rectangles in the upper-left quadrant of the tree-map. Concatenation of the panel and cell-lines fields present in the GI50 data set formed the names that appear in the property list to the left of the tree-map. The selected property corresponds to the OVCAR-3 cell line from the OVA panel. The data for the OVA-OVCAR-3 property was truncated to the range 4.0-8.0 within the MPX software. Compounds with an OVA-OVCAR-3 below 4.0 were assigned the value 4.0, and compounds with a above 8.0 were assigned the value 8.0. The legend to the right of the tree-map indicates that 1689 of the 1883 compounds have a GIs0 between 4 and 5,128 compounds have a between 5 and 6,36 compounds have a between 6 and 7, and 14 compounds have a between 7 and 8. In addition, there are 16 compounds with unreported Compound identification and property value are displayed in a tool-tip as user passes the mouse over the tree-map.

The data set consists of a variety of diverse compound classes, clustered by chemical structure. Consequently, there are multiple regions of SAR scattered throughout the tree-map of Figure 2. For example, compounds belonging to the camptothecin and ellipticine classes, both of which are known potent inhibitors of tumor growth, as discussed in Ohashi, M.; Oki, T. Ellipticine and related anticancer agents. Expert Opin. Ther. Pat. 1996, 6, 1285-1294 and Shi, L. M.; Myers, T. G.; Fan, Y.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the National Cancer Institute Anticancer Drug Discovery Database: Cluster Analysis of Ellipticine Analogs with p53-Inverse and Central Nervous System-Selective Patterns of Activity. Molecular Pharmacology 1998, 53, 241 - 251, are located within grid A-6 of the tree-map. In the event one was interested in visualizing all compounds active against the OVCAR-3, OVCAR-4, OVCAR-5, and OVCAR-8 cell lines independent of their chemical class. Such visualization is achieved in MPX by clustering the data on the basis of a property profile as illustrated in Figure 3. The OVCAR-3, OVCAR-4, OVCAR-5, and OVCAR-8 cell lines were selected from the property list, and the data was re-clustered (employing Euclidean distance and complete linkage) using the tool bar's cluster button. Two distinct regions of densely packed rectangles characterize the treemap of the re-clustered data. These regions correspond to compounds that are either potent or impotent inhibitors of tumor cell growth across the selected cell lines. Compounds with intermediate profiles separate these two regions. The most potent inhibitors of tumor cell growth are located within grids A-2 and 8-2 of the tree-map.

A tree-map can represent only one property at a time, whereby a heatmap allows visualization of multiple properties simultaneously. A heatmap of the GI50 data set clustered by profile across the OVCAR-3, OVCAR-4, OVCAR-5, and OVCAR-8 cell lines is shown in Figure 4. The potent inhibitors within grids A-2 and B-2 of the tree-map of Figure 3 are represented in the upper region of the heatmap. Compound identification, property name and property value are displayed in a tool-tip as user passes the mouse over the heatmap. Eight compounds within this region have been selected and the structures of the first four of these are shown in the Structure Visualizer at the bottom of Figure 4. The compounds within this region are structurally diverse as indicated by a centroid of 0.5450 computed for the eight compounds selected. The compound identified by NSC 176327 is the current selection in the Structure Visualizer, and its corresponding location in the heatmap is highlighted in red. When the user toggles the display back to the tree-map, the corresponding tree-map rectangle for this

compound is likewise highlighted in red. The same color is a common compound identifier in the tree-map and heatmap.

Generally, prior art commercial software applications that provide support for heatmaps do so only in the context of two-way clustering. That is, the criteria used to cluster across rows of the heatmap must be used to cluster across the columns. Heatmaps created within MPX allow two different sets of criteria to be used to cluster across row and columns. The advantages gained from treating rows and columns of the heatmap as independent clusters are illustrated in the analysis of the NCTRER data set that follows.

The NCTRER data set consists of 230 compounds from a variety of chemical classes. The data set includes the properties "Activity Category ER-RBA" and "ChemClass ERB". Activity Category ER-RBA classifies the estrogen receptor binding strength of each compound as one of the following: inactive, slight binder, active weak, active medium, and active strong. ChemClass ERB assigns each compound to one of six broad chemical classes: miscellaneous, biphenyls, diethylstilbestrol (DES), diphenlymethanes, phenols, phytoestrogens, and steroids. Sub-types are used to further define compounds within these classes. For example, phytoestrogen compounds fall into one of the following sub-classes: flavones, isoflavones, and mycoestrogens. The MPX software is compatible with both continuous numeric and categorical text data. Hence, the text values associated with Activity Category ER-RBA and ChemClass ERB properties did not have to be numerically encoded prior to analysis.

The compounds in the NCTRER data set were clustered by 2D chemical structure employing I-Tanimoto distance and complete linkage. The heatmap of the clustered compounds and the clustered properties Activity Category ER-RBA and ChemClass ERB is shown in Figure 5. The group average similarity for these compounds is 0.5788, and the balance of the hierarchical cluster of compounds is 0.3478. Compounds assigned to the "inactive" Activity Category are colored yellow, and "active strong" compounds are colored light blue. Assignments within ChemClass ERB also are represented in the heatmap by rectangles filled with various shades of yellow or blue. Relationships between chemical structure and estrogen binding receptor activity are readily apparent from the heatmap of Figure 5. The chemical classes with the highest estrogen receptor binding affinity are identified on the right hand side of the heatmap. It is important to note the correlation between the ordering of these compounds in the hierarchical cluster and their assigned ChemClass ERB. Also, the most active compounds (light blue) within each class lie adjacent to one another in the heatmap.

The NCTRER data set was partitioned into a smaller subset on the basis of Activity Category being active medium or active strong. The compounds in the subset were clustered by 2D chemical structure, which resulted in a general organization of the compounds by chemical class as is evident from the tree-map of Figure 6. ChemClass ERB assignment is used to shade each rectangle of the tree-map, and regions occupied by compounds belonging to the various chemical classes have been added to the figure. The large rectangle occupying the right half of the tree-map corresponds to the compound kepone, an unusual estrogen receptor binder assigned to the chemical class "miscellaneous" and structurally dissimilar to every other compound in the subset.

The structures of four phytoestrogen isoflavones from grid A-I, and four steroids from grid C-I of the tree-map in Figure 6 are shown in Figure 7A and 7B, respectively. The interpretation of the SAR within the two sets of structures is straightforward. Isoflavones become potent binders to estrogenic receptors when hydroxyl groups in the 7 and 4' positions mimic 4, 4' OH positions in diethylstilbestrol,

as illustrated by genistein. Steroids possessing 3-hydroxy substituted, phenolic, A-rings bind to estrogen receptors, and the strength of this binding is increased when an oxygen atom is present at the 17-position.

In order to create a tree-map or heatmap from structure-property data, one must first generate a hierarchical cluster of the chemical structures. There are a number of techniques for clustering chemical structures, and agglomerative hierarchical clustering using the reciprocal nearest neighbors (RNN) algorithm such as disclosed in Murtagh, F. A Survey of Recent Advances in Hierarchical Clustering Algorithms. Computer Journal 1983,26(4), 354-359, represents one approach. The RNN algorithm is performed in two distinct steps.

The first step involves computing a two-dimensional matrix of the distance between all pairs of chemical structures in the data set. There are a number of ways of computing the distance between two chemical structures, and one approach is to compute the distance as I-Tanimoto. The Tanimoto coefficient computed from the binary representations for a pair of chemical structures obtained from a computation of their chemical fingerprints, and the following equation (noted: this is the equation noted earlier and is not renumbered for that reason): where: a is the number of 1 bits appearing in the binary representation of structure A, b is the number of 1 bits appearing in the binary representation of structure B, and c is the number of 1 bits in common to the binary representations of both structures. A number of algorithms exist for computing chemical fingerprints, including the Daylight algorithm such as disclosed in . The guide to Daylight theory is available online: http:/www.daylight.com/dayhtml/doc/theo~/theory.finger.html.

One second step of the RNN algorithm involves linking the individual chemical structures together to form the cluster hierarchy. Ward's method, as disclosed in Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. J. Amer. Stat. Assoc. 1963, 58, 236-244, is a common approach for the linking phase of the RNN algorithm.

A hierarchical cluster of chemical structures is depicted as a tree-map by a recursive algorithm described as follows. The tree-map algorithm traverses the branches of a hierarchical cluster recursively beginning with the root node. The algorithm begins with a rectangular region corresponding to the root of the hierarchical cluster. As each branch is visited, a rectangular region of the treemap is split evenly along alternating vertical and horizontal centers. Upon reaching a terminal node in the hierarchical cluster, the corresponding rectangular region of the tree-map is associated with the terminal node of the cluster. The previously discussed pseudo-code illustrates one possible implementation of the tree-map algorithm.

With specific reference to Figures 2-7B, there is shown one embodiment of the invention. The properties associated with the chemical structures represented in the tree-map are shown in the list on the left side of Figure 2. The tree-map occupies the central portion of the figure, and the colors used in the tree-map to represent a chemical structure's value for a specific property are defined in the legend to the right of the tree-map. The numbers to the right of the colored rectangles in the legend indicate the number of rectangular regions assigned that color in the tree-map. The tree-map consists of regions of densely packed rectangles that are separated by more sparse regions. The dense regions represent sub-clusters of structurally related compounds. The names of compounds, their structures and property values are presented in a tool-tip as the mouse cursor passes over the various rectangular regions in the tree-map.

In one further preferred embodiment, a selected rectangle or defined region is recursively partitioned to provide second level tree-map (not shown) for cognitive visualization and analysis.

The data in Figure 2 were re-clustered based on the compound's values across the four specified properties (OVA-OVCAR-3, OVA-OVCAR-4, OVAOVCAR- 5, and OVA-OVCAR-8) and the resulting tree-map is shown in the drawings.

The heatmap of the cluster of compounds and the cluster of selected properties (OVA-OVCAR-3,OVA-OVCAR-4,OVA-OVCAR-5, and OVA-OVCAR- 8) is shown in the drawings. As the mouse cursor passes over the heatmap image, a tool-tip displays the name of the compound, the property and the property value. The heatmap provides visualization each compound's response across a range of selected properties. Each column of the heatmap corresponds to one of the four selected properties. The rows of the heatmap correspond to the order in which compounds appear in the hierarchical cluster of chemical structures.

Schneiderman Tree-Maps

With specific reference to Figures 8-10, there are shown prior art treemaps as disclosed and shown in Baehrecke, E. H., Dang, N., Babaria, K., and Shneiderman, B.; "Visualization and analysis of microarray and gene ontology data with treemaps," BMC Bioinformatics, 2004, 5(1), 84-96 (the "Shneiderman tree-maps"). The Shneiderman tree-maps relate to the results of microarray gene expression experiments mapped onto the gene ontology and do not relate to specific chemical structure-property data. The Shneiderman tree-maps are based on a predefined hierarchy, namely, the gene ontology. That is, the gene ontology defines the division of the tree-map into specific rectangular regions, and then the gene expression data is mapped onto these fixed regions. The rectangular regions of the tree-map serve merely as boundaries between elements of the gene ontology, and do not represent relationships between the elements of the gene expression data being depicted. The Shneiderman treemaps necessarily are replete with explanatory verbiage related to the gene ontology, rendering them not readily visually cognitive.

In marked contrast to the Shneiderman tree-maps, the layout or configuration of the tree-map of the present invention depends on the hierarchy that results from clustering the specific chemical structure-property data. That is, the division of a tree-map into rectangular regions is dynamic (it will be different for every data set and clustering criteria employed), and the size and arrangement of these rectangular regions conveys specific information about the relationship between chemical structures in the cluster hierarchy. An advantage of representing a hierarchical cluster of chemical structures as a tree-map is that the tree-map allows one to simultaneously visualize all sub-clusters within the hierarchy.

MPX Applications

The MPX method and system provides a number of applications. The MPX method may be used qualitatively to predict the properties of new compounds. A button on the tool bar provides mechanism for adding new chemical structures to a data set. As new compounds are added, the data set is re-clustered and the tree-map is redrawn to reflect the placement of the new compounds within the cluster. Assuming a sufficient number of compounds defining an SAR or SPR exists within the original data set, the activity of new compounds may be inferred from their nearest neighbors within the cluster. One may assess the appropriateness of such comparisons by comparing the structures of nearest neighbors in a Structure Visualizer dialog. If a new compound appears centered within a dense

region of the tree-map with a defined SAR or SPR and the Structure Visualizer confirms the new compound is structurally similar to its nearest neighbors, then the properties of those neighbors may be ascribed to the new compound.

Further, the combination of hierarchical clustering based on 20 chemical structure and visualization as a tree-map provides a novel approach to representing the topology of the chemical space for a set of chemical structures. Properties other than those relating to biological activity may be mapped onto this topology. (See the definition and examples of "property" as set out hereinbefore). For example, the date on which a compound was synthesized and the name of the corresponding therapeutic program could be incorporated into the data set. Such information would allow one to visualize the discovery process within and across therapeutic projects. Shading the rectangles of the tree-map by date provides a historic representation of the various medicinal chemistry strategies applied within a project. A tree-map encoding the name of the therapeutic project for which a compound was synthesized might be used to identify compounds applicable to other therapeutic programs.

Another useful property within the contemplation of the present invention is the visualization of chemical compound of related patent information (e.g. assignee, inventor(s), claims relating the chemical structures under analysis). For example, one using the MPX method could color code the assignee, search the patent information and determine which commonly owned patents disclose andlor claim certain chemical structures.

With specific reference to Figure 11, there is shown an embodiment of this contemplation. The data presented in this figure was obtained from a patent search of GABA-A a5 agonists. The tree-map of Figure 11 represents the clustering of the GABA-A a5 agonists based on the similarity of their chemical structures, and the rectangular regions of the tree-map are colored according to the assignee of the patented chemical matter.

The development of software capable of representing multidimensional structure-property data in a straightforward and intuitive manner is a challenge achieved by the present invention. The simultaneous representation of clustered data as heatmaps and tree-maps is a dynamic means of visualizing SAR or SPR. That is, these combined two powerful visualizations in combination with a set of data-mining tools into a software application provides a dynamic method and system for exploring and understanding multidimensional, structure-activity data sets. The MPX method may be used to identify regions of structural similarity and dissimilarity within a data of compounds, segregate compounds into distinct regions on the basis of a defined activity profile, and visualize relationships between structure and activity.

The MPX system is preferably applied to moderately sized data sets of up to about 10,000 compounds, and most preferably, about 5,000 to 8,000 compounds. Larger data sets may take significantly longer to cluster, and might then not be well represented as heatmaps and tree-maps.

The MPX method and system may be applied to a broad range of object and related subject data, and is not limited to chemical structure (object) and chemical properties (subjects) data. Any object-subjects data compatible with hierarchical clustering, is within the contemplation of the present invention. The subjects data is related to the object data, and the object data is organized into a hierarchical cluster.

While the foregoing description and examples used rectangles as the defined regions of the tree-map, it is within the contemplation of the present invention to use other geometric regions (e.g.

triangles, hexagons, and the like).It is further within the contemplation of the present invention to use combinations of geometric regions for display (e.g., for example rectangles, triangles, and hexagons in combination) and optionally varying the selection of the geometric shape depending upon a particular type of object-subject data or structure-. Is it also within the contemplation of the present invention to employ a different selection of geometric regions depending upon hierarchical position, iterative analysis step or other selection decision.

It will also be recognized by those of appropriate skill in the art that the present invention is not limited to the display of chemical related structures, but in effect, is applicable to a wide array of data assemblies capable of any rational hierarchical structure. For example, medical data (e.g., heart and liver treatment data), therapeutic intervention data (treatment of a specific disorder), genealogical data, commercial and investing data, market analysis and penetration data, voting and polling data, and other forms of linkable data.

It is also contemplated that beyond an initial broad data display, additional recursive portioning or focusing of a user-selected region may occur to provide a second, tertiary or greater level viewing with the tree-map and heat map. It is additionally contemplated that three-dimensional viewing of the data is available.

This invention may be embodied in other specific forms without departing from the essential characteristics as described herein. The embodiments described above are to be considered in all respects as illustrative only and not restrictive in any manner. The scope of the invention is indicated by the following claims rather than by the foregoing description. Any and all changes which come within the meaning and range of equivalency of the claims are to be considered within their scope.

CLAIMS

WHAT IS CLAIMED IS:

1. A method for the readily cognitive display of object data for a plurality of related subjects or structure property data of a plurality of compounds, comprising the steps of:

(a) displaying the data in a tree-map;

(b) displaying the data in a heatmap; and

(c) integrating the tree-map and heatmap, wherein there is a readily cognitive display of the data.

2. The method of claim 1, wherein the data is hierarchically structurable.

3. The method of claim 1, further comprising organizing the data in a cluster hierarchy for display in the tree-map and heatmap.

4. The method of claim 1, wherein the tree-map comprises a plurality of defined regions, and the heatmap comprises a plurality of cells comprising row and column intersections.

5. The method of claim 4, wherein there is a 1 :l correspondence between a specific heatmap cell and a specific tree-map region.

6. The method of claim 1, further comprising distinguishing an identifier for each subject, and wherein the distinguishing identifier for each subject is the same in the heatmap and the tree-map.

7. The method of claim 1, wherein the tree-map comprises a plurality of defined regions, and said method further comprises recursive nesting within a selected defined region to provide one of a second order defined region and a record order derivative tree-map.

8. The method of claim 1, wherein the plurality of related subject data comprise chemical compound structures and the object data comprises properties the plurality of compounds.

9. The method of claim 1, wherein the plurality of related subjects comprise chemical structures and the object data are informational materials related to the chemical structures.

10. The method of claim 9, wherein the informational materials comprise patent information.

11. The method of claim 1, wherein the property comprises a physico-chemical property related to the plurality of compounds.

12. The method of claim 11, wherein the compounds have a related structure.

13. The method of claim 1, further comprising a step of organizing the data, wherein the step of organizing comprises applying a reciprocal nearest neighbor algorithm to the data to generate a cluster hierarchy.

14. A method for the cognitive display of structure-property data for a plurality of chemical compounds comprising:

(a) organizing the structure-property data by agglomerative hierarchical clustering; and

(b) displaying the structure-property data in a tree-map; wherein there is a readily cognitive display of the structure-property data.

15. The method of claim 14, further comprising a step of creating a first structure-property data set and a second structure property data set, and respective tree-maps for the data sets, whereby said tree-maps comprise defined regions, and wherein the regions are differently sized for the respective data sets.

16. The method of claim 15, wherein the defined regions comprising geometric shapes enabling a ready two-dimensional display, said geometric shapes selected from the group consisting of rectangles, squares, triangles, hexagons, and combinations of the same.

17. The method of claim 14, wherein step (a) further comprises a step of applying a reciprocal nearest neighbor algorithm to the data to generate the hierarchical clustering.

18. A method for the readily cognitive visual display of structure-property data for a plurality of chemical compounds, comprising:

(a) organizing the structure-property data by agglomerative hierarchical clustering;

(b) displaying the structure-property data in a tree-map;

(c) displaying the structure-property data in a heatmap; and (d) integrating the tree-map and heatmap, whereby there is a readily cognitive multidimensional display of the structure-property data.

FIG. 1B
Hierarchical cluster of ten compounds as a heatmap

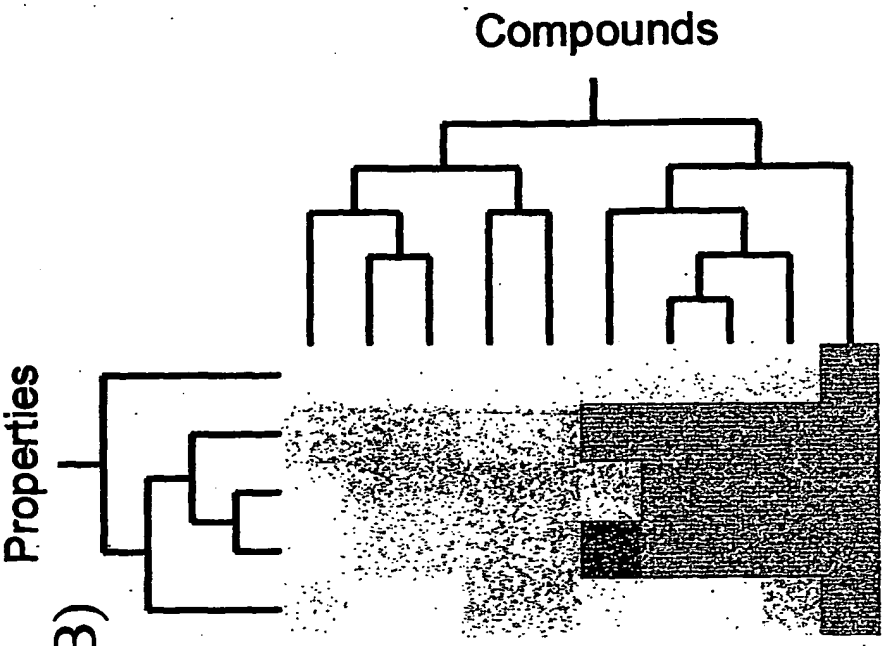
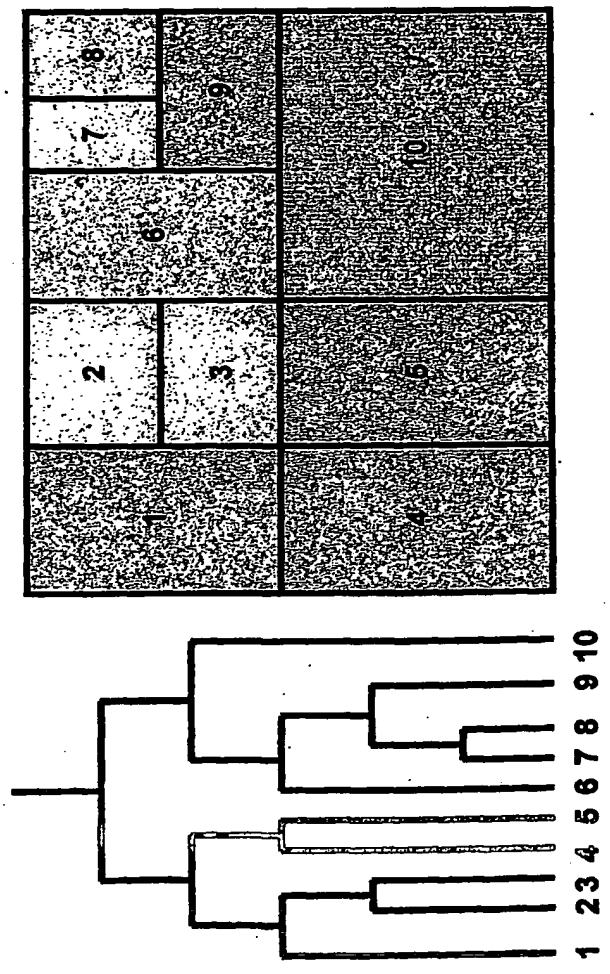


FIG. 1A
Hierarchical cluster of ten compounds as a tree-map



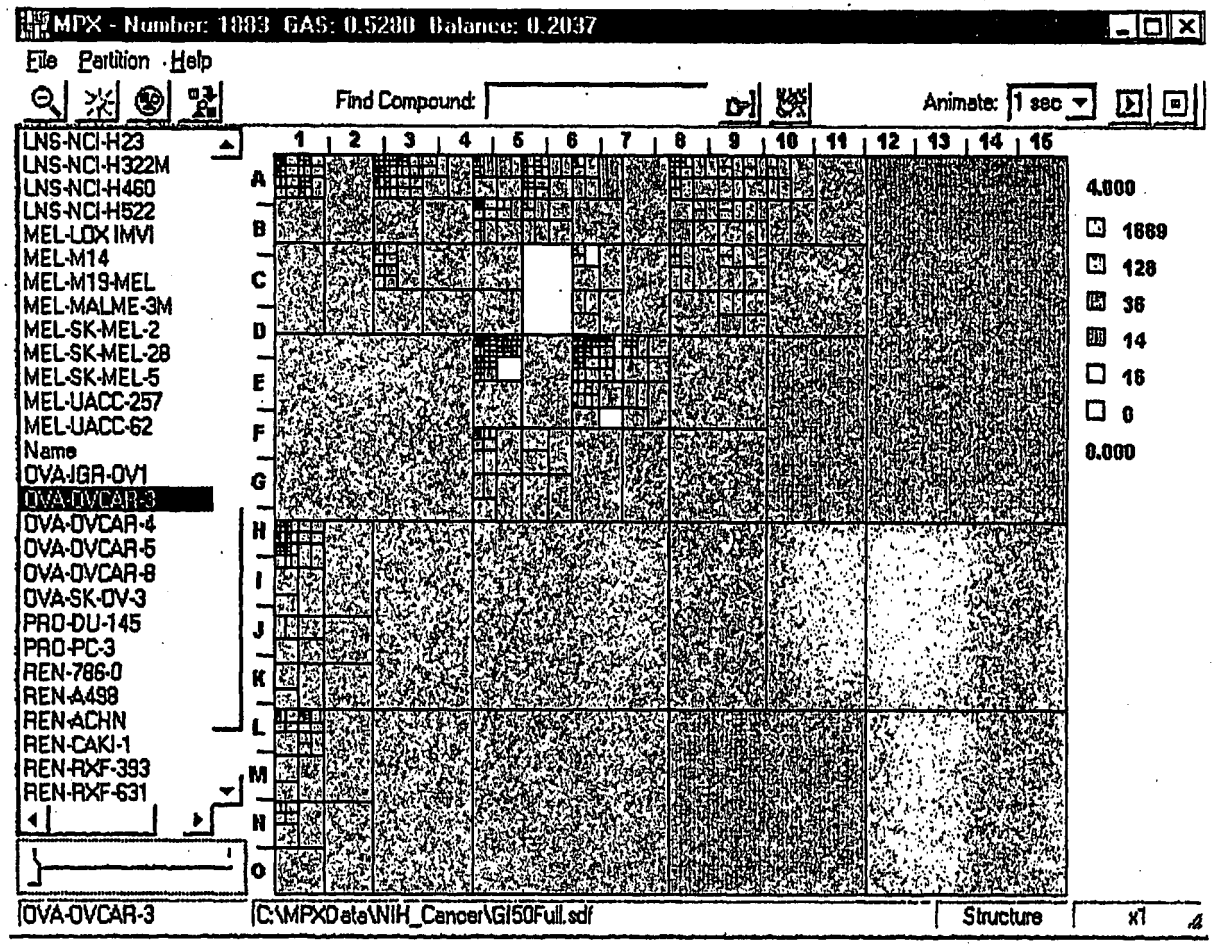


FIG. 2

A tree-map of 1883 compounds from the NCI GI_{50} diversity data set. The compounds are clustered according to the similarity between their 2D chemical fingerprints. The rectangles of the tree-map are colored according to each compound's GI_{50} in the OVCAR-3 cell line.

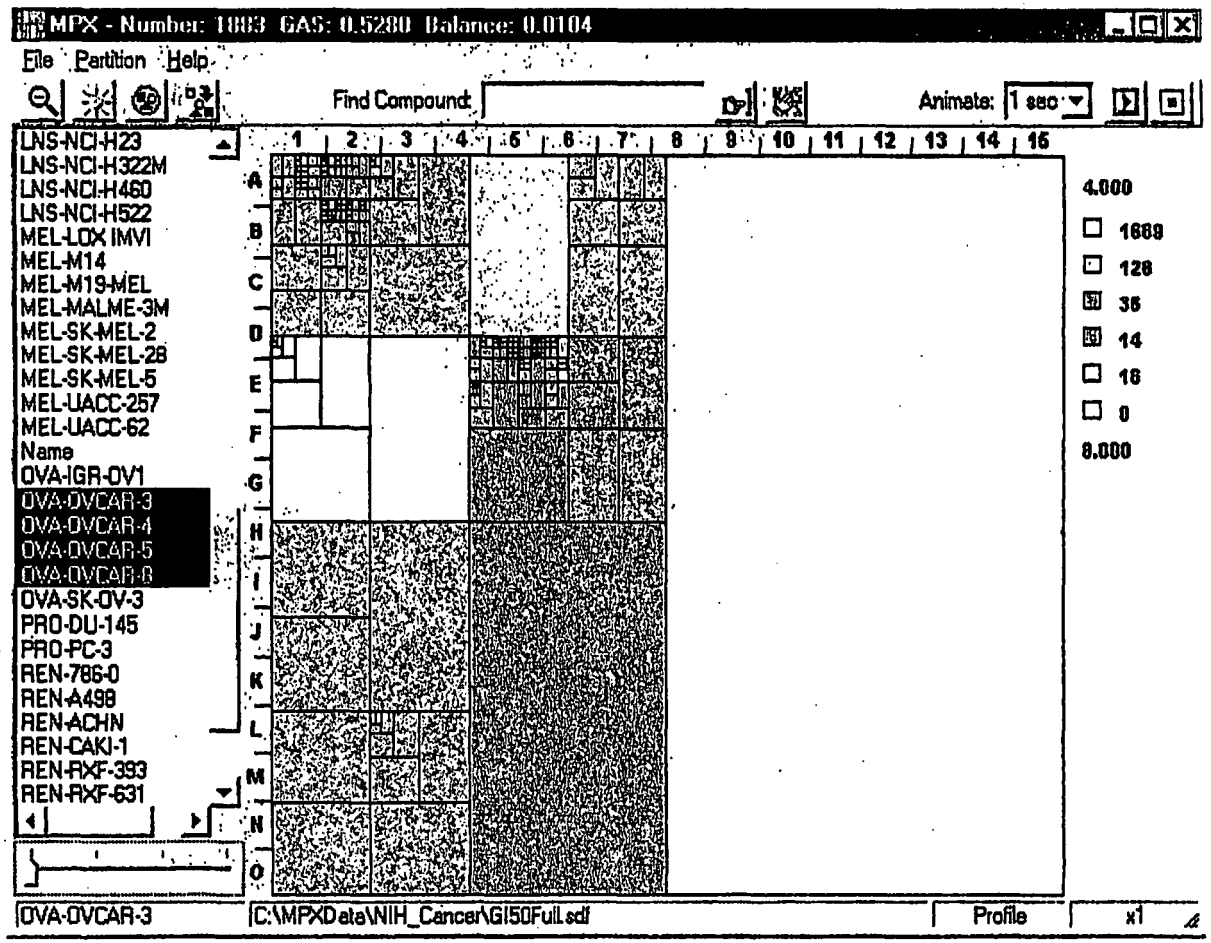


FIG. 3

A tree-map of 1883 compounds from the NCI GI_{50} diversity data set. The compounds are clustered on the basis of their GI_{50} profile across the OVCAR-3, -4, -5, and -8 cell lines. The rectangles of the tree-map are shaded according to each compound's GI_{50} in the OVCAR-3 cell line.

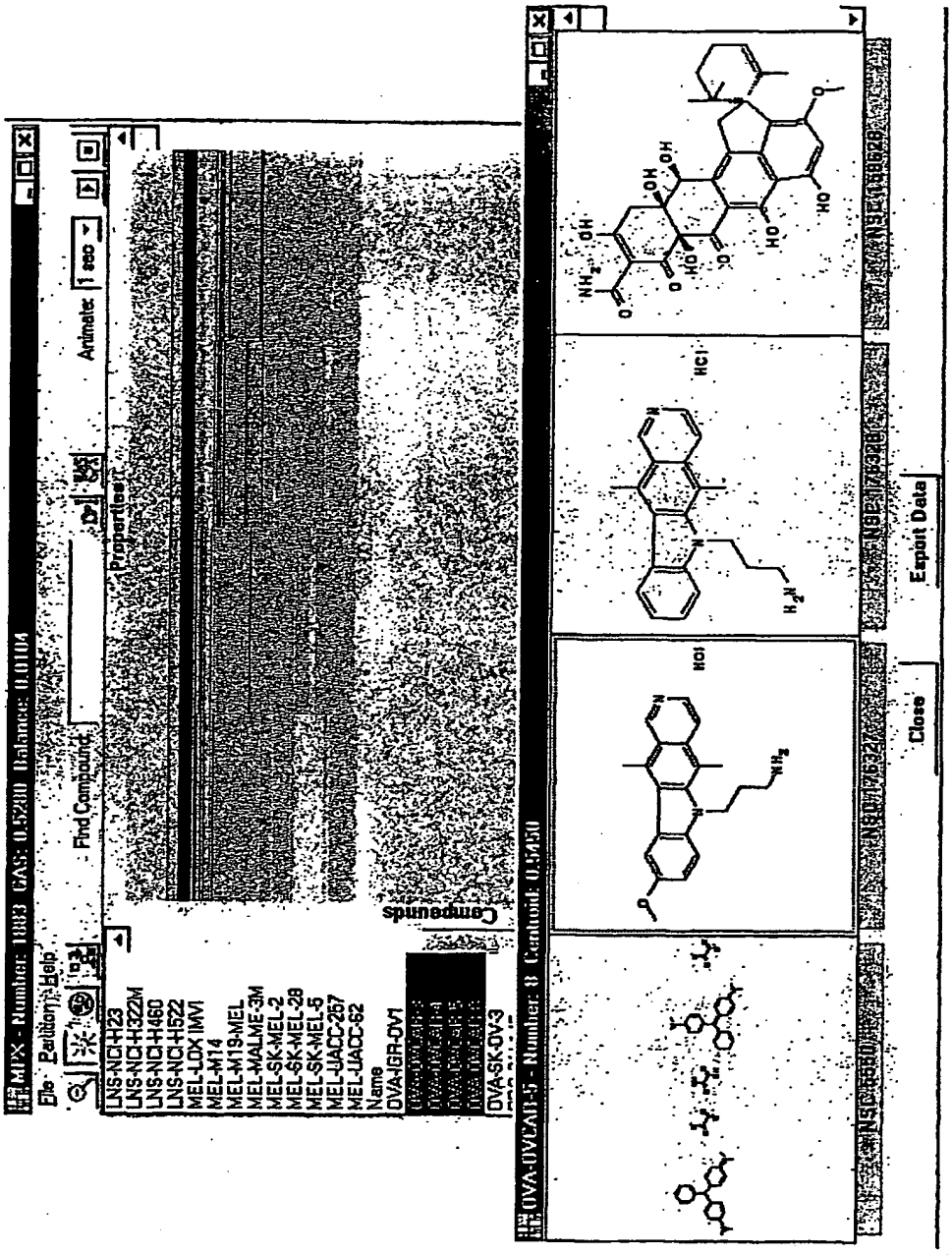


FIG. 4

The heatmap representation of the clustered profile from Figure 3. Compounds with high GI_{50} values across the OVCAR-3, -4, -5, and -8 cell lines are visible in the upper half of the heatmap. A portion of this region of the heatmap has been selected, and the corresponding structures for these compounds are shown in the dialog below the heatmap.

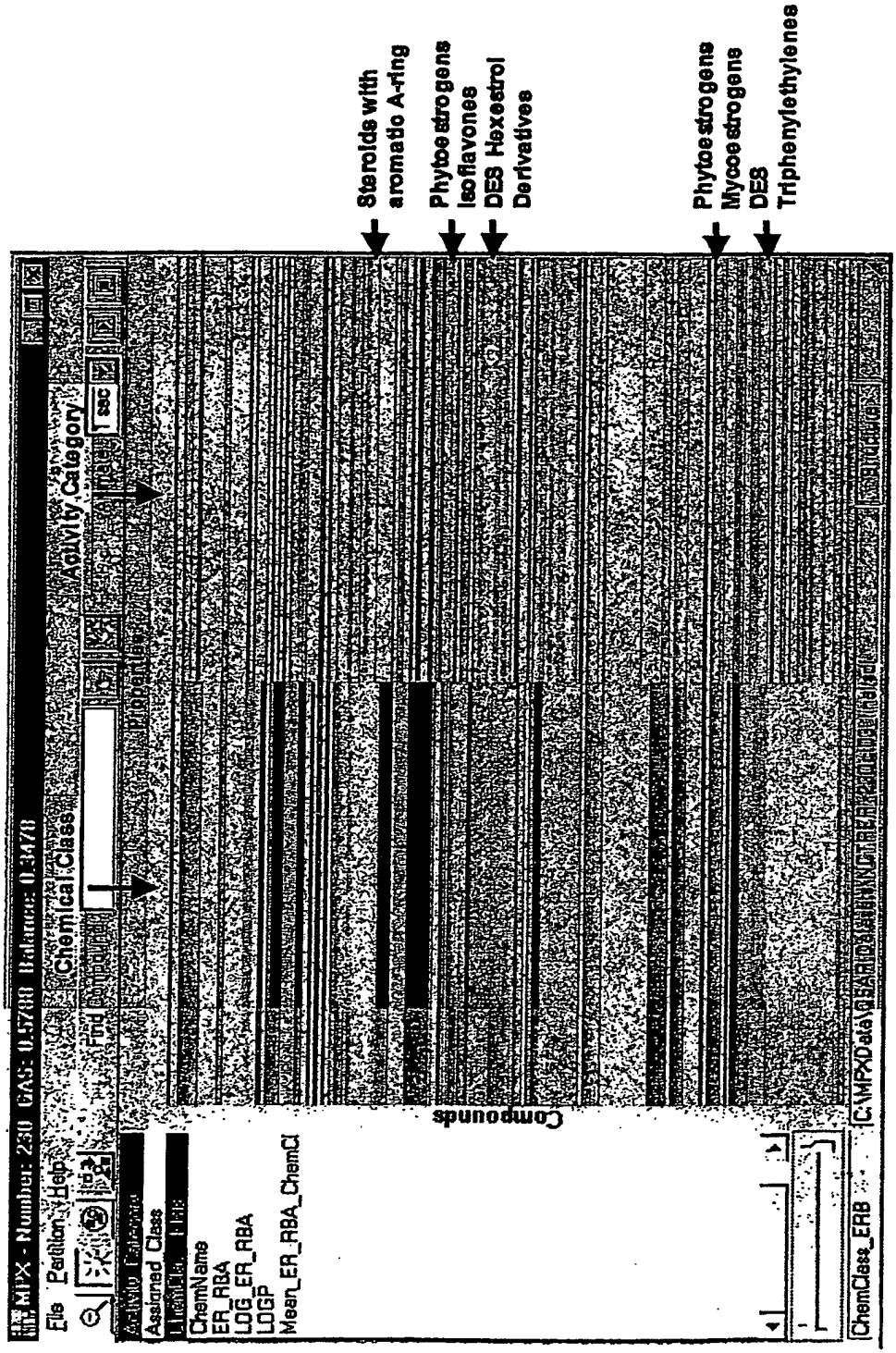


FIG. 5

A heatmap of the NCTRER estrogen receptor binding data set. The compounds in the data set are clustered according to the similarity between their 2D chemical fingerprints. A relationship between structure and activity is apparent from the overlap of activity category and assigned chemical class for the most active estrogen receptor binding compounds in the data set

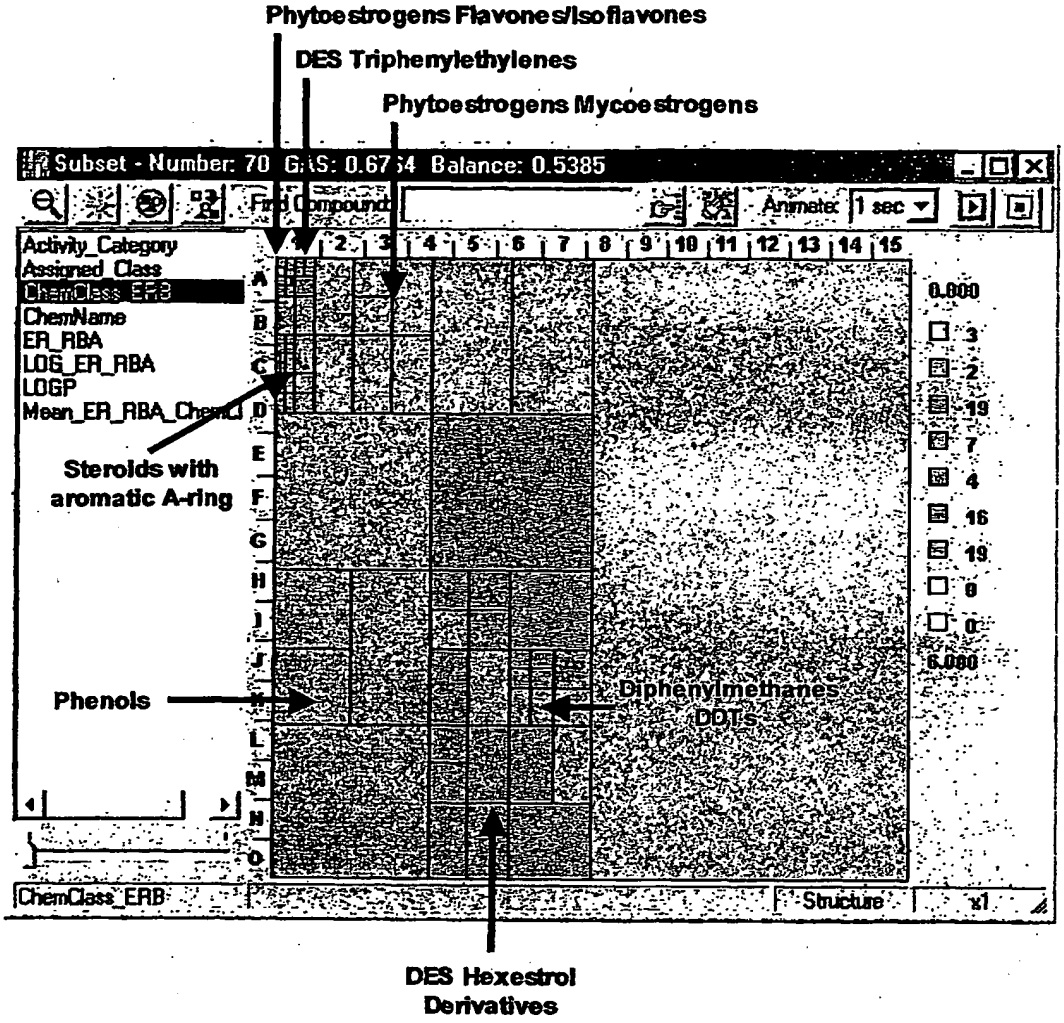


FIG. 6

A tree-map of the most active estrogen receptor binding compounds in the NCTRER data set. The compounds are clustered according to the similarity between their 2D chemical fingerprints. The compounds tend to cluster according to their assigned chemical class as illustrated by the annotations that have been added to the tree-map.

FIG. 7A

Structures of genistein and three related phytoestrogen isoflavones selected from the tree-map of Figure 6

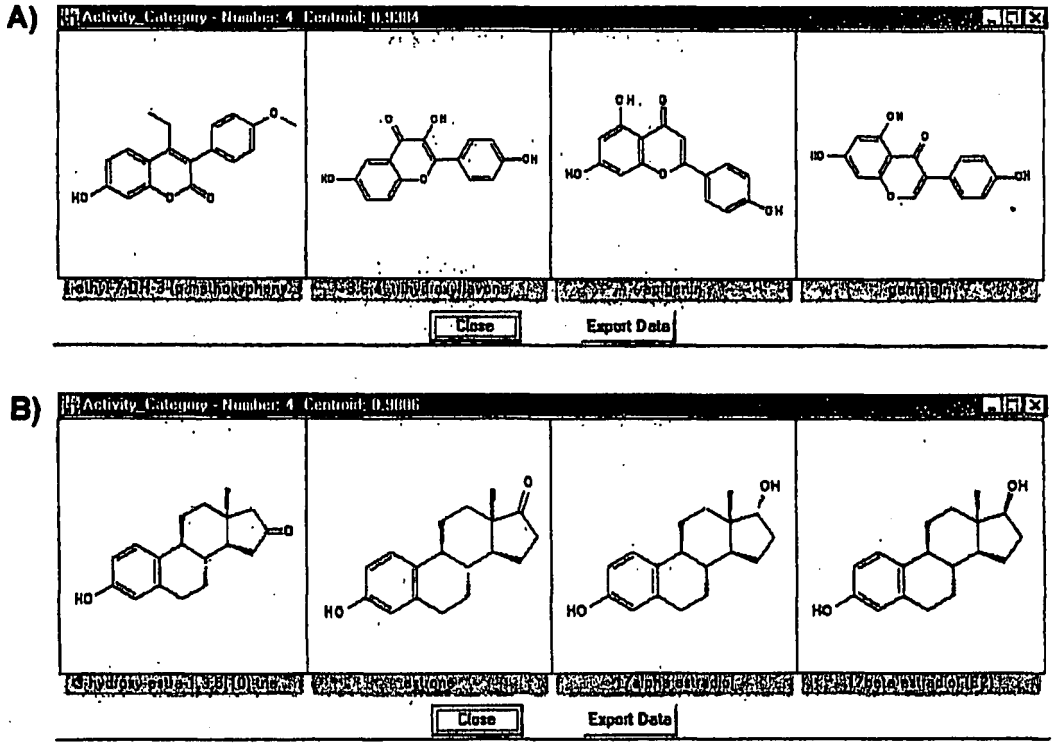


FIG. 7B

Four steroids possessing aromatic A-rings selected from the tree-map of Figure 6.

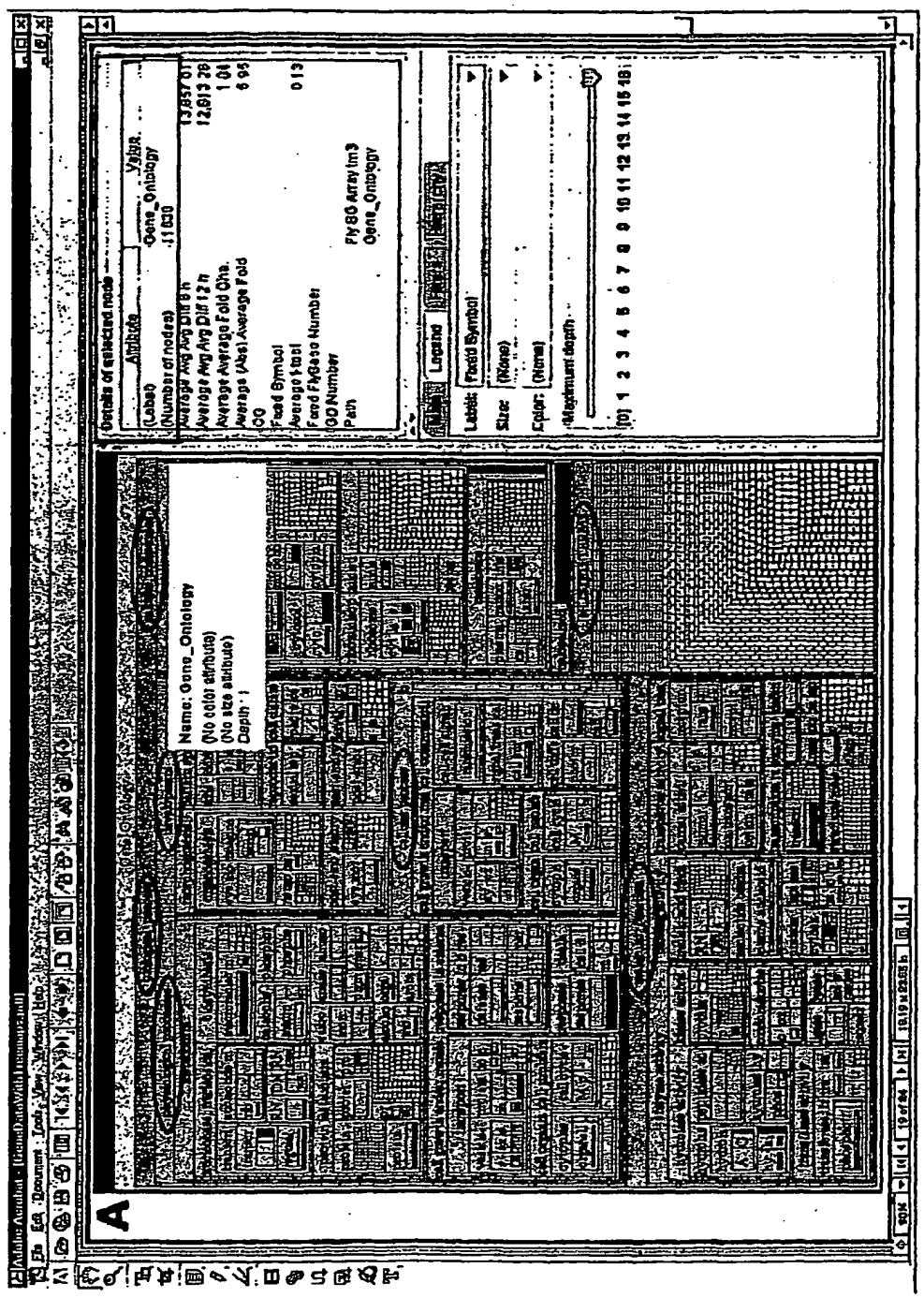


FIG. 8

Tree-map of the entire fly genome in the context of the Gene Ontology

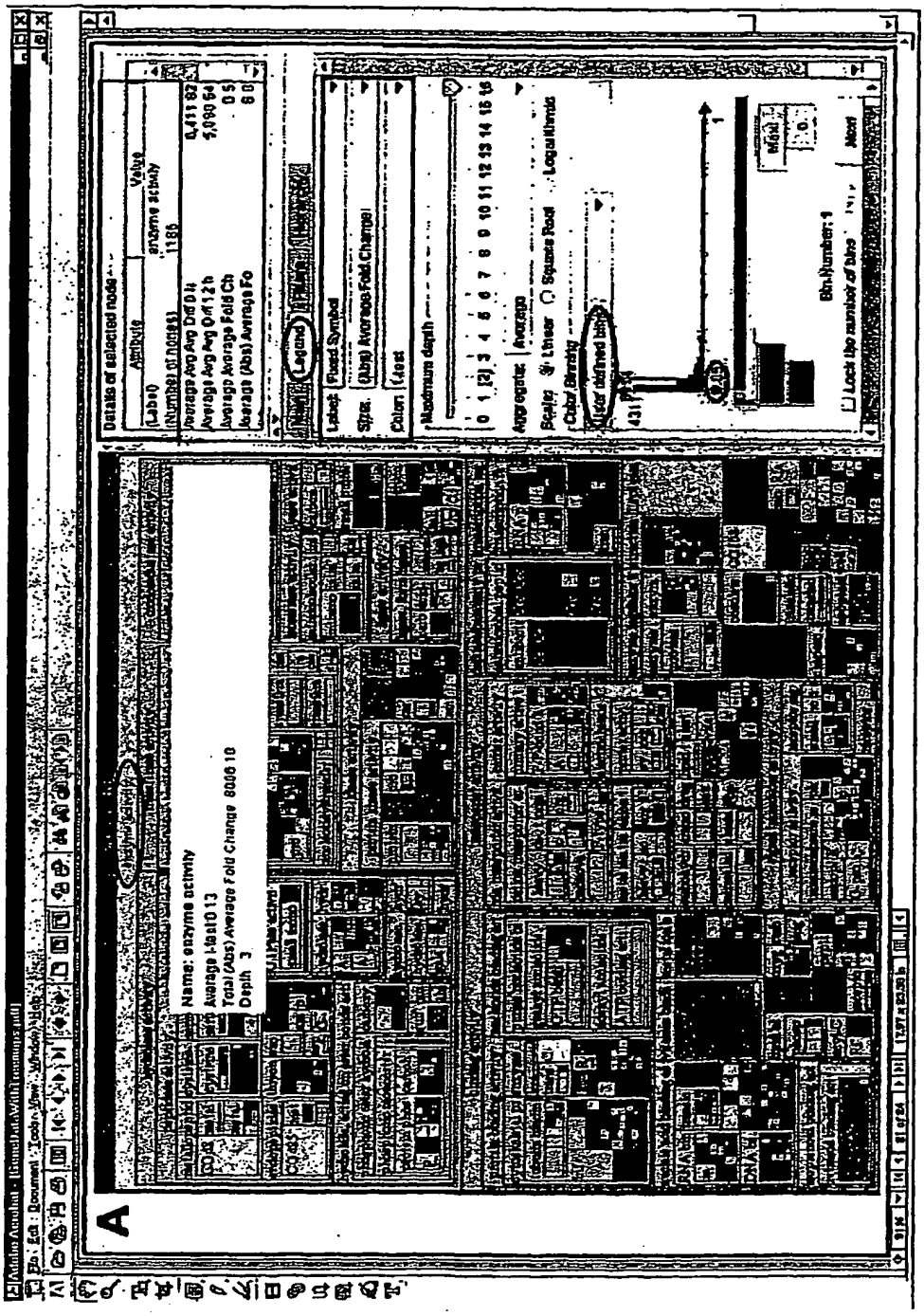


FIG. 9

Selection of the enzyme_activity region of the molecular_function region from the tree-map of Fig. 8

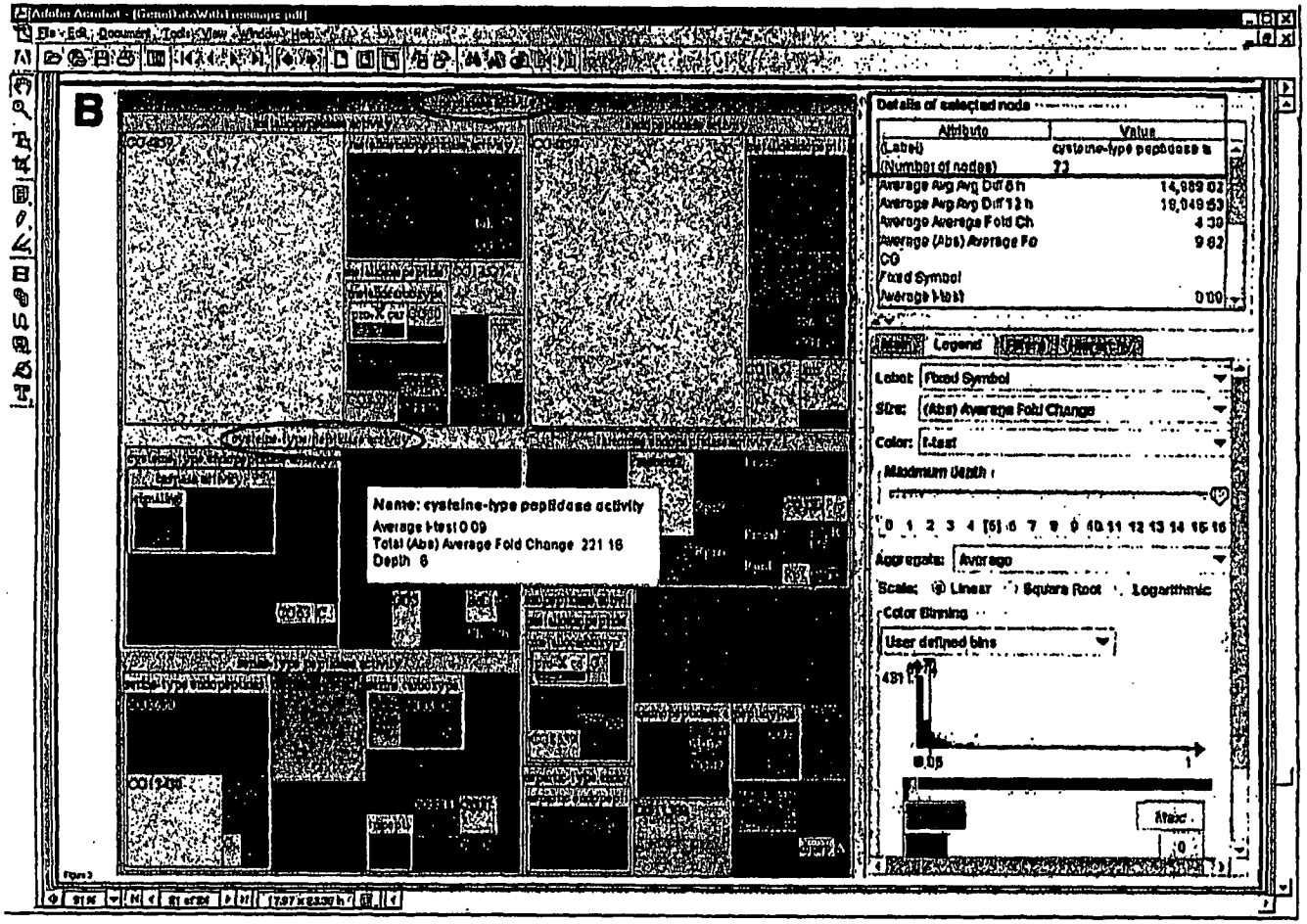


FIG. 10

Selection of the peptidase_activity region from Fig. 9 showing which genes of the fly genome are associated with various peptidase enzymes. The size and color of the rectangles displayed are based on the parameters set in the legend panel to the right of the tree-map.

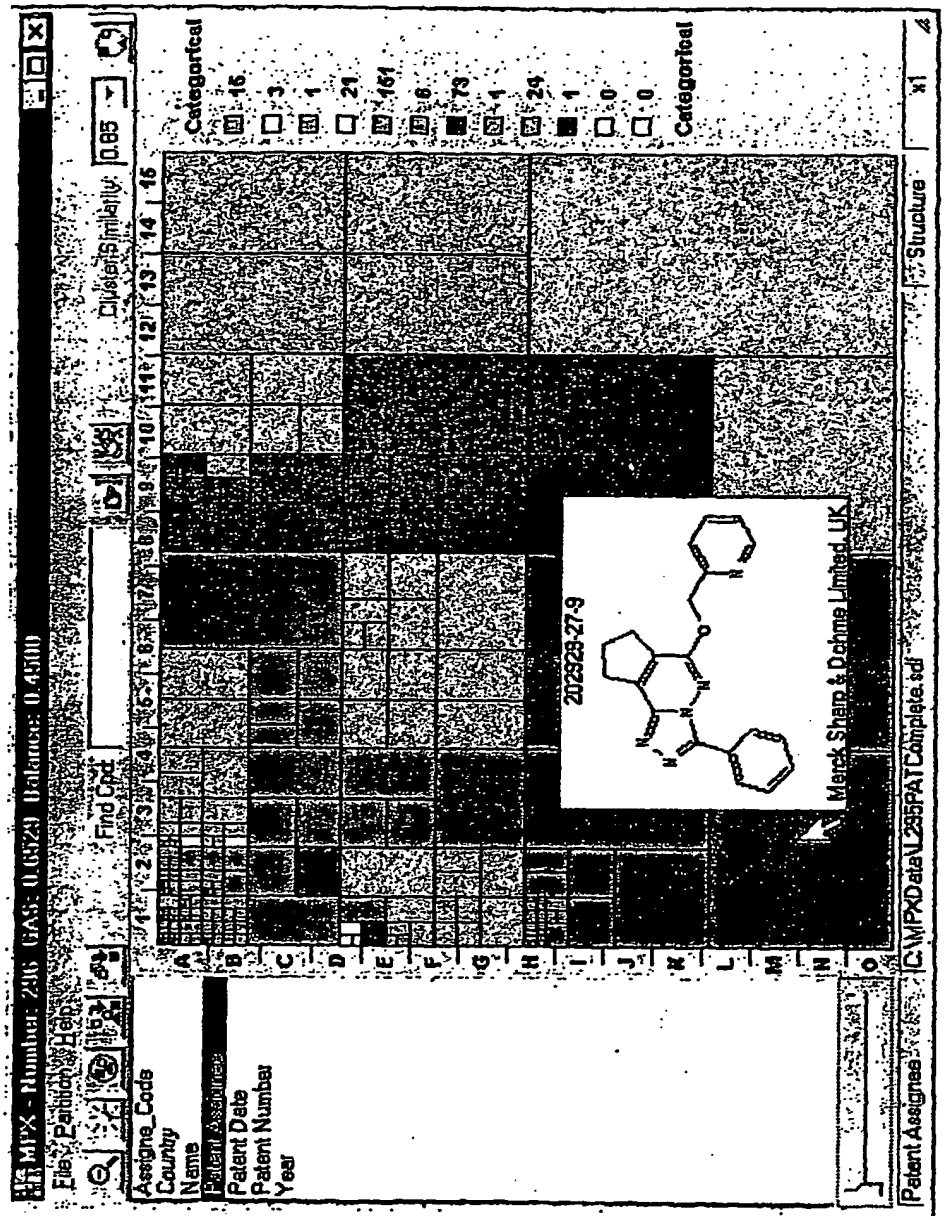

MPX - Number: 286 GAS: 0.6529 Balance: 0.4500
File Partition Help
Find Cpd
Diversity Similarity 0.85
Assigne_Code
Country
Name
Patent Assignee
Patent Date
Patent Number
Year
Categorical
16
3
1
21
161
6
73
1
24
1
0
0
Categorical
202928-27-9
Merck Sharp & Dohme Limited,UK
Patent Assignee
C:\MPXData\L295PATComplete.sdf
Structure
x1

Figure 11

This Page is inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☑ BLURED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLORED OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REPERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning documents *will not* correct images problems checked, please do not report the problems to the IFW Image Problem Mailbox